Execution Time Prediction for Energy-Efficient Hardware Accelerators

Tao Chen, Alex Rucker, and G. Edward Suh
Computer Systems Laboratory
Cornell University

Accelerators in Interactive Computing Systems

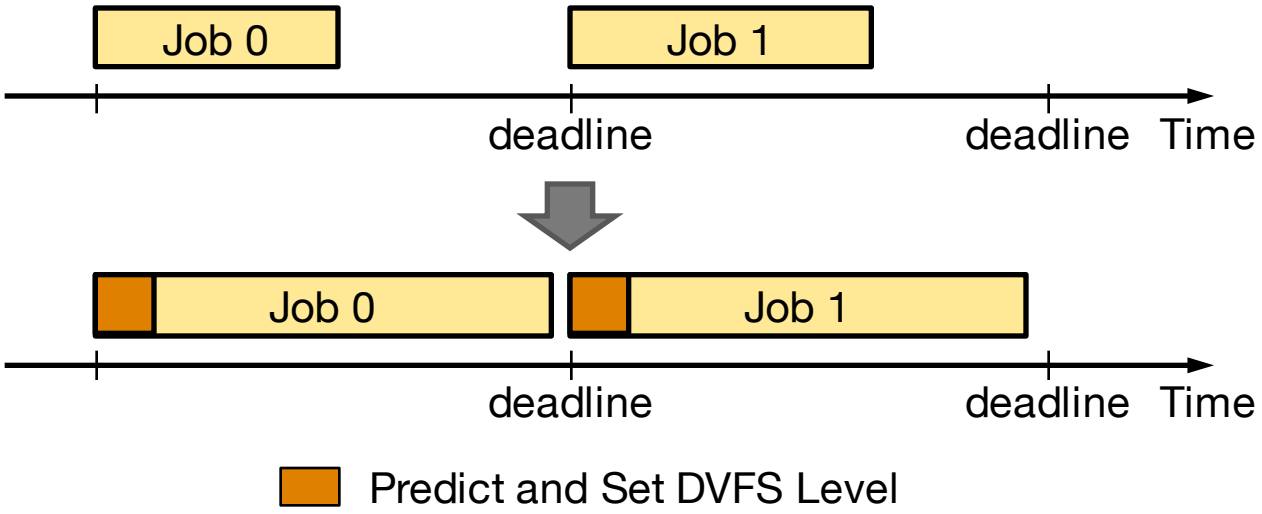
- Interactive systems have response time requirements and often use hardware accelerators

- **Observation**: Finishing earlier than the requirement is usually not needed

- **Goal**: Perform DVFS for hardware accelerators to save energy while meeting response time requirements

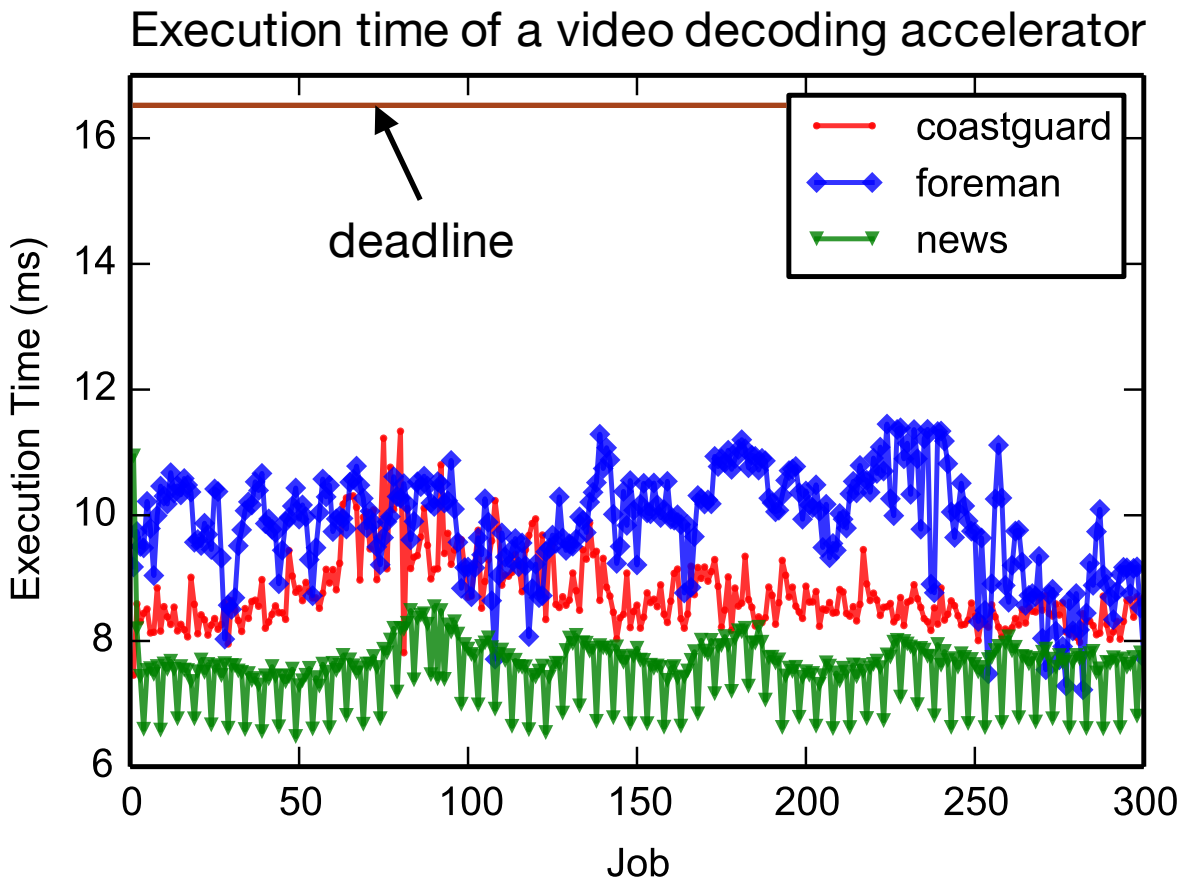
DVFS for Interactive Computing Systems

- Save energy by running slower (lower frequency/voltage)



- Requirement
  - Correctly predict each job's execution time

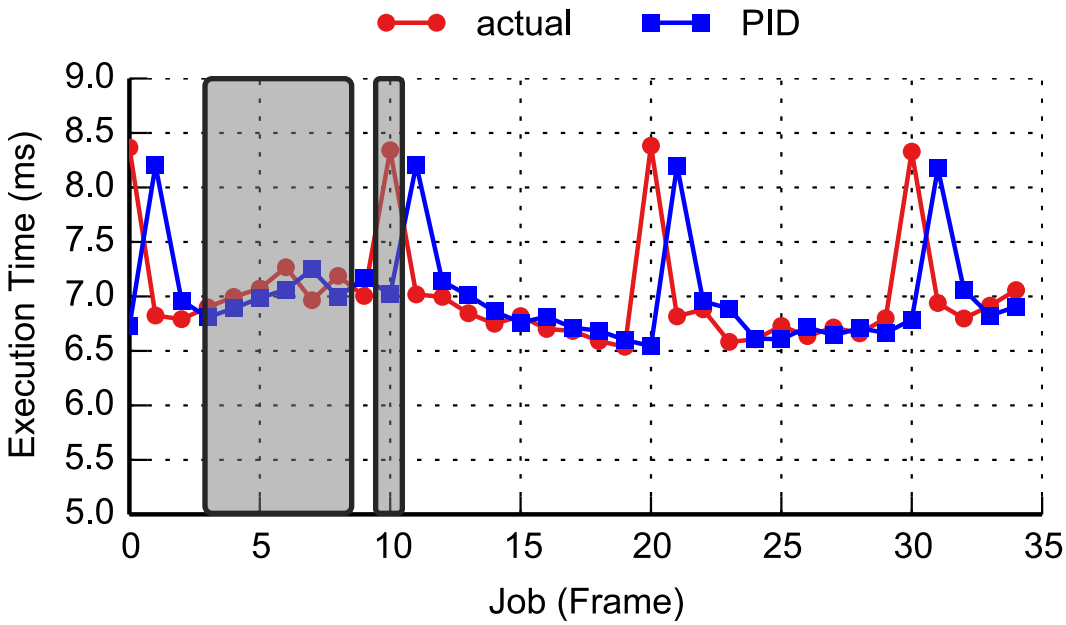
Opportunity and Challenge



- Opportunity: Most jobs finish earlier than the deadline
- Challenge: Irregular variations in job execution time

Conventional DVFS Controllers

- History-based execution time prediction
  - Example: PID controller



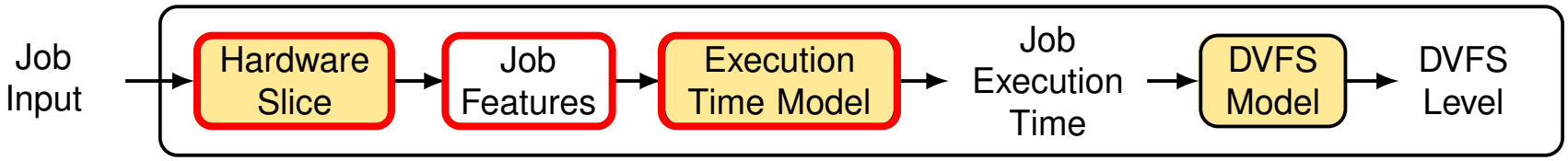
- Problem of history-based prediction
  - Reactive — decisions lag behind changes

Predictive DVFS Framework for Accelerators

- Approach: Build a predictor hardware for each accelerator that uses job input data to predict execution time

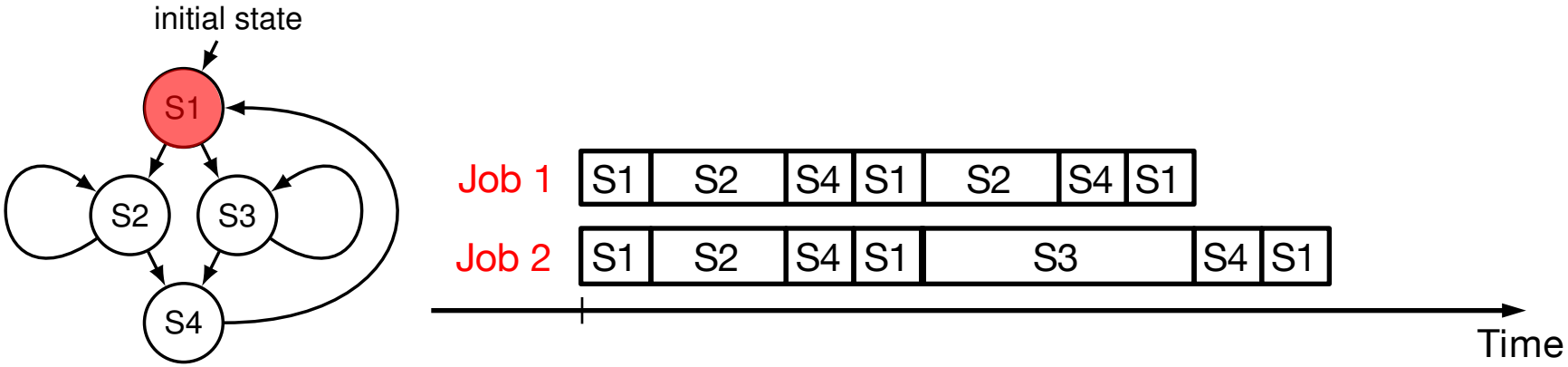
- Design Time: Build predictor and train prediction model
  - Identify features related to execution time
  - Generate a hardware slice that can calculate features quickly
  - Train a prediction model that maps features to execution time

- Run Time: Run predictor to inform DVFS decisions



Features to Capture Execution Time Variation

- **Source of variation**: input-dependent control decisions



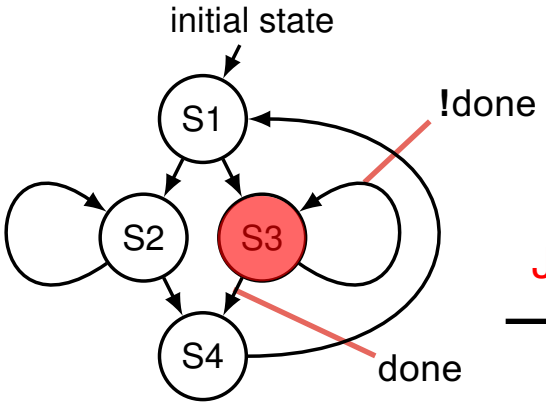
- **Feature**: State Transition Count

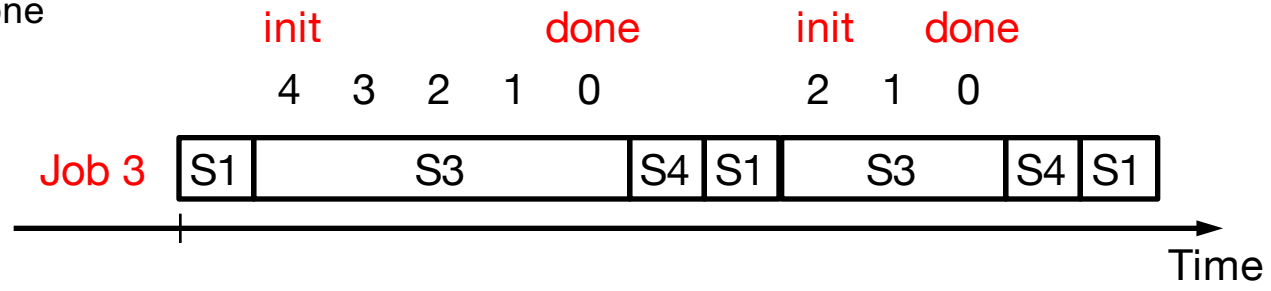
$$STC = [st_{1,2}, st_{1,3}, st_{2,4}, st_{3,4}, st_{4,1}]$$

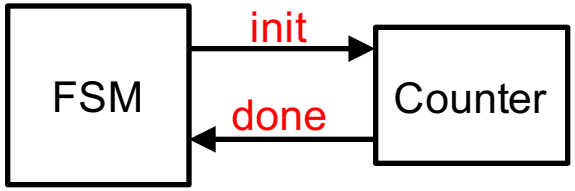
Job 1	2	0	2	0	2
Job 2	1	1	1	1	2

Features to Capture Execution Time Variation

- Variable state latency







- **Feature**: Counter Average Initial Value

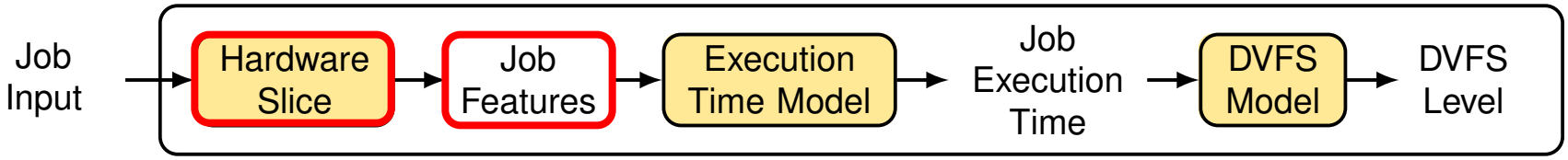
$$AIV = [iv_{S3}]$$

Job 3 3

- Other counter features in the paper

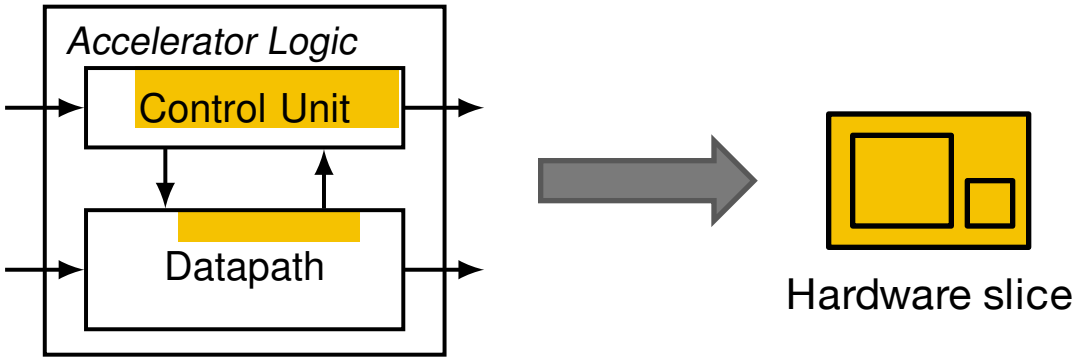
Identifying and Extracting Features

- Automated flow based on RTL analysis
  - Identify FSM and counter features in RTL
  - Instrument RTL to extract features
- More details in the paper

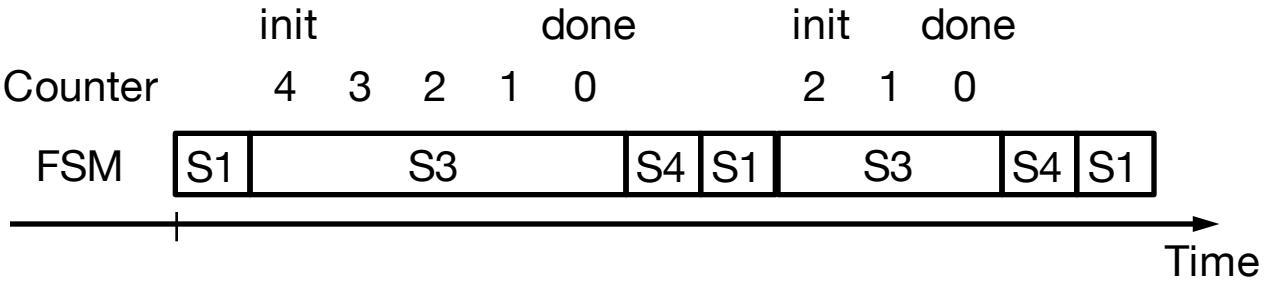


Hardware Slicing

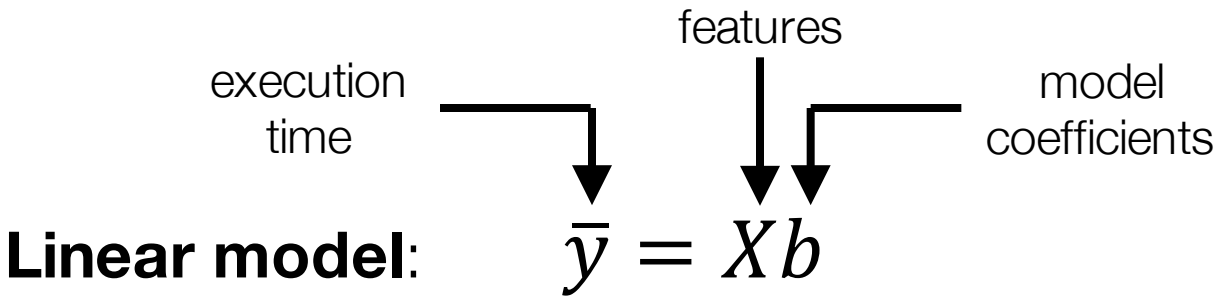
- Need to obtain features before running the accelerator
- Create a minimal version of the accelerator
  - Program slicing on accelerator RTL code



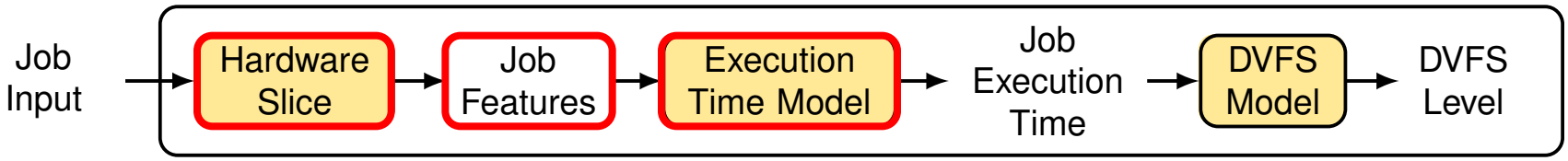
- Optimize hardware slice to run fast



Execution Time Prediction Model



- Train model using convex optimization
  - Reduce the number of features
  - Prioritize meeting deadlines over saving energy



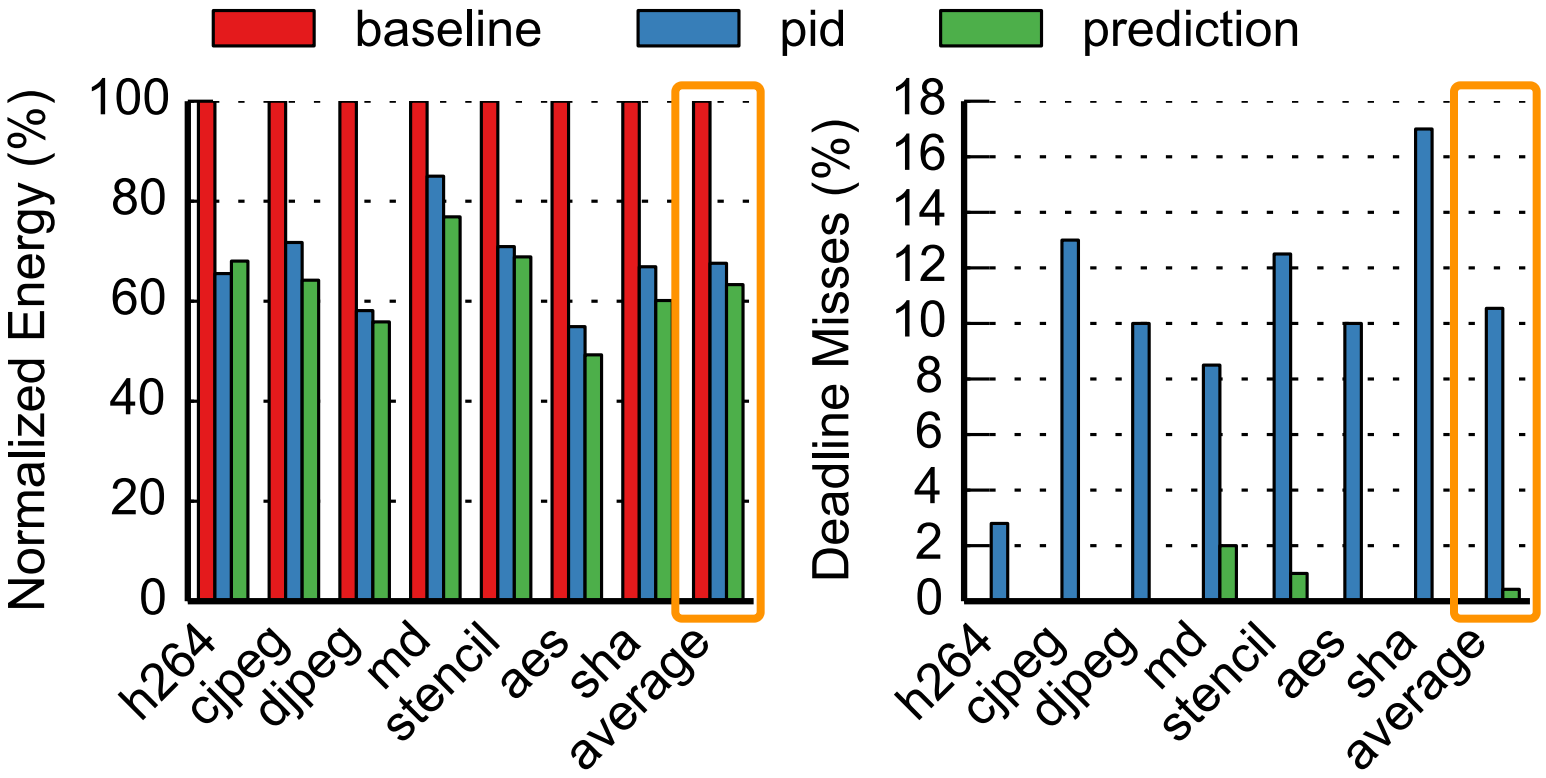
Evaluation Methodology

- Vertically integrated evaluation methodology
  - Circuit-level simulation: obtain voltage-frequency relationship
  - Gate-level modeling: obtain area, power and energy numbers
  - Register-transfer-level simulation: obtain execution time
- Benchmark accelerators

Name	Description
h264	Video decoding
cjpeg	Image encoding
djpeg	Image decoding
aes	Cryptography
sha	Cryptography
md	Molecular dynamics
stencil	Image processing

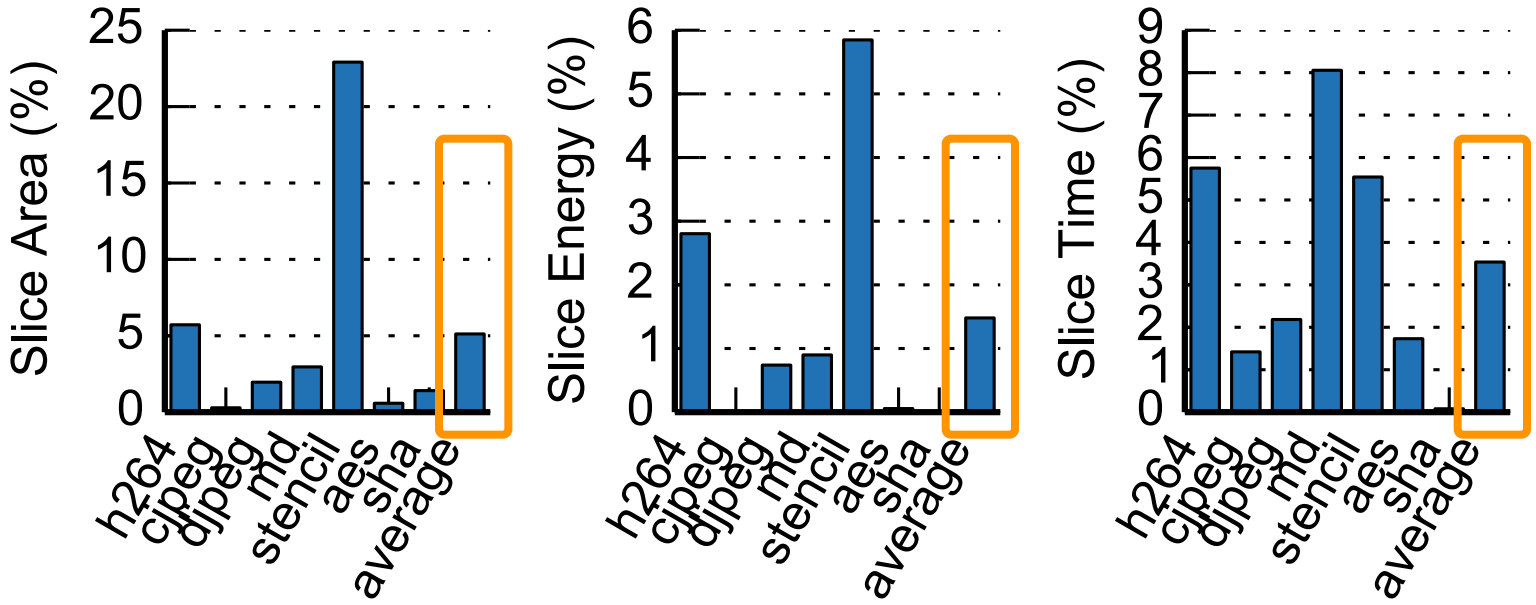
- Deadline: 16.7 ms

Results: Energy and Deadline Misses



- **36.7%** energy savings on average
- **0.4%** deadline misses

Results: Overheads of Slice-Based Predictor



- **5.1%** area overhead
- **1.5%** energy overhead
- **3.5%** execution time overhead

More Evaluation Results in Paper

- More detailed experimental results
  - Prediction Accuracy Analysis
  - Results with Predictor Overheads Removed
  - Sensitivity Study on Varying Deadlines

- Platform extensions
  - DVFS with Voltage Boosting
  - Results for FPGA-based Accelerators
  - Results for Accelerators Generated by HLS

Summary

Observation: Finishing faster than the deadline is not needed

Goal: DVFS for accelerators with response time requirements

Solution: Prediction-based DVFS

- Execution time depends on input-dependent control decisions
- Hardware features can be used to capture control decisions
- Proposed a framework to generate predictors automatically

Results: Highly accurate DVFS for accelerators

Questions?

Execution Time Prediction for Energy-Efficient Hardware Accelerators

Tao Chen, Alex Rucker, and G. Edward Suh
Computer Systems Laboratory
Cornell University